Sentiment Analysis on Arabic Reviews

Basant Allam, John Sedrak, Lojain El Sayed

May 8, 2023

1 Introduction

Sentiment analysis is a rapidly growing field of research that aims to automatically extract and classify opinions, attitudes, and emotions expressed in text data. With the increase and spread of digitalized information and social media, sentiment analysis has become a vital tool for businesses, governments, and researchers to collect opinions on products, services, and events. However, most of the existing sentiment analysis research has focused on English and other widely spoken languages, leaving a gap in the analysis of sentiment in languages with a smaller digital presence. Arabic, being the fifth most spoken language in the world, is one of the languages that has received much less attention in the field of sentiment analysis.

This report aims to explore sentiment analysis on a set of hotel, book, and airline reviews and classify them as positive, negative, or mixed while tackling challenges such as spelling mistakes and lack of diacritic usage. The report also aims to highlight the challenges of performing sentiment analysis on Arabic text.

2 Challenges and Approach

Performing sentiment analysis on any data collected from the internet introduces several challenges such as URLs, hashtags, misspelled words, slang words, and words with repeated letters (eg. Wooow amazinggg!!). For Arabic text specifically, there is the added issue of handling the lack of diacritics known as "Tashkeel" (تشكيل) and the lower availability of data in comparison with English.

2.1 Misspellings and Slang

Misspellings, slang, and words with repeated letters are by far the most common challenge that is faced when performing any sort of natural language processing on internet-collected data, such as reviews and tweets, because there are no regulations or restrictions that require maintaining proper grammar or spelling.

A reasonable approach would be to process all the words and remove letters that appear more than two times in succession (ie. thrice in a row). We allow for double letters to include words such as "color" (اللون). We can then use a spellchecker to find the closest candidate for every word, which will eliminate (to a good extent) misspellings and any invalid double letters. We are considering two possible approaches for slang words:

1. Creating a dictionary for common slang words and a non-slang word to replace each with.
2. Leaving them as is and allow the fine tuning to learn their meanings.

We will try both approaches and determine which yields better results.

2.2 URLs and Hashtags

URLs and hashtags are a natural consequence of collecting any web based data. In the context of sentiment analysis, the content of URL strings is seldom useful. Hashtags are usually several words concatenated into a single word, which makes them gibberish to the sentiment analysis model, so they should not be included as well.

A reasonable approach would be to replace all URLs and hashtags with a [URL] token and a [HASHTAG] token respectively. This way, we remove gibberish while maintaining context in the corpus.

2.3 Diacritics

Figure 1: A sample Arabic text. Please note that Harakat is another word for Tashkeel [3].

Arabic is one of the four abjad languages. Abjad languages have scripts that only write the consonants and long vowels as letters, leaving the short vowels to be handled by diacritics (as shown in Figure 1). In Arabic, these diacritics are known as tashkeel (تشكيل). For this reason, many words have the same spelling, but differ in their diacritics. Usually, native Arabic speakers can easily infer the diacritics on a word given its context, and thus infer the meaning of the word itself. For this reason, diacritics are usually omitted from text, making it difficult for computers to determine which word embedding should be used when.

We are considering two potential approaches for dealing with diacritics:

1. Using diacritic restoration tools to estimate diacritics from contexts.
2. Removing diacritics from all words to somewhat simplify the task.

We will attempt both approaches and determine which approach yields better results.

2.4 Availability

While Arabic is a very rich language with numerous pieces of literature and poetry, the amount of user-generated Arabic content on the internet is very limited. Furthermore, spoken Arabic and formal, written Arabic are widely different, so models cannot be solely trained on written Arabic then applied directly to casual spoken Arabic. When dealing with internet reviews and tweets, most people write the same way they would speak. Therefore, in order to perform sentiment analysis on internet content, models should be trained – or at the very least fine-tuned – on internet content.

To overcome the challenge of limited data availability, we will be using a pre-trained model and fine-tuning it on our dataset.

3 Pre-processing

The preprocessing steps we decided to have are:

1. Remove Diacritics
2. Remove Punctuation Marks
3. Tokenization
4. Padding

We decided not to remove stop words nor do lemmatization as the model we will use is BERT so lemmatization is unnecessary. We also found that the dataset has no URLs nor hashtags thus there was no need to remove them in pre-processing.

4 Dataset

The dataset we will be using is 100K Arabic Reviews Dataset[2]. This is an Arabic Dataset consisting of 99,999 reviews written in Arabic and it is mainly used in Sentiment Analysis tasks. The reviews are about hotels, books and airlines. The Hotel and book reviews from the HARD[4] and BRAD [1] datasets respectively, while the airline reviews were collected manually from 100 airlines reviews. The dataset has 2 fields: the review and its label being either positive, negative or mixed.

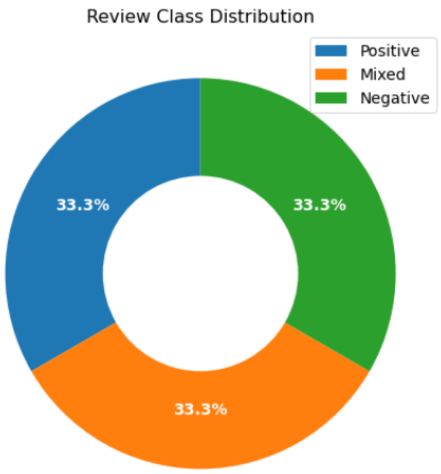
Class	Sample Data Points
Positive	ممتاز . كل شي ممتاز في الفندق طاقم الاستقبال رائع والفندق رائع ونظيف.
Mixed	مرضي. سعر الغرف غالي جدا
Negative	سوء أدب موظف الاستقبال مخيب للأمل.

Figure 2: Sample Data from 100K Reviews Dataset. [2].

4.1 Data Analysis

In order to better understand the dataset, we did some data analysis to find certain aspects of the dataset such as class distribution, length, spelling mistakes.

4.1.1 Class Distribution



Class	Data Points
Positive	33,333
Negative	33,333
Mixed	33,333

Figure 3: Class Distribution of Dataset.

The data distribution is balanced having an equal number of datapoints in each of the three classes, which is promising as it lowers the risk of bias while training the model. As shown in Figure 3.

4.1.2 Length of Reviews

Although it is generally more common that people write more when writing a negative review, compared to a positive review, this is not the case with our dataset. We found that the word count

of reviews in all three classes are almost the same with negligible differences, as shown in Figure 4. Moreover, for all 3 classes the mean character counts are also very close as shown in Figure 4.

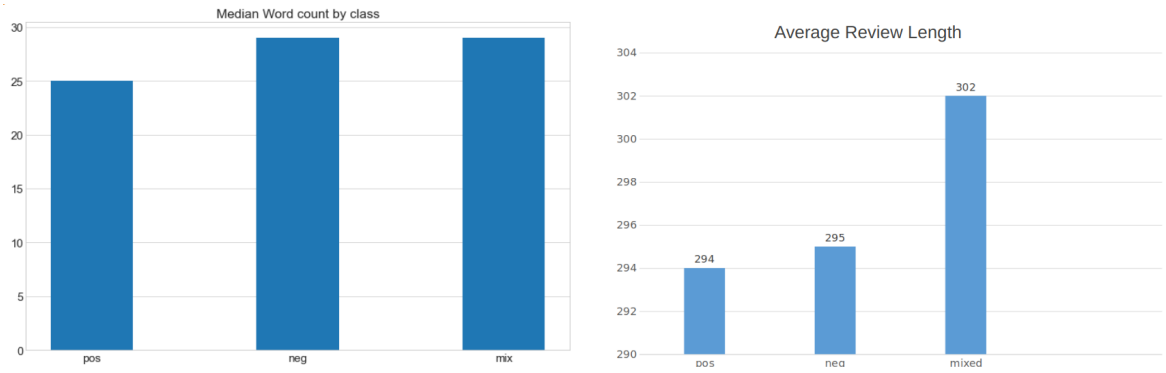


Figure 4: The left figure shows the median word count of every review class. The right figure shows mean review length in terms of number of characters.

4.1.3 Spelling Mistakes

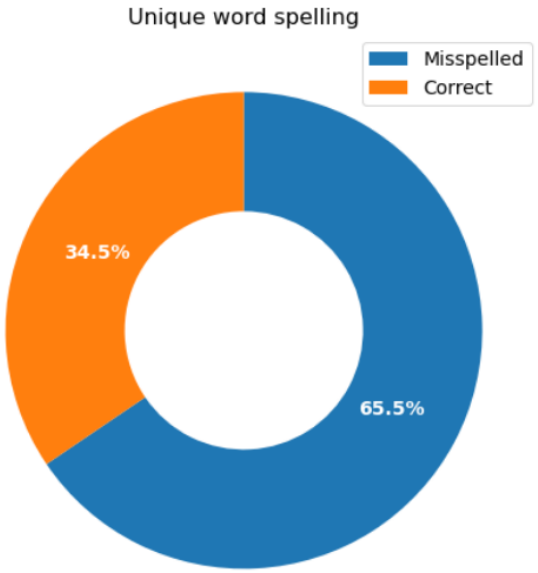


Figure 5: The number of unique words that were misspelled before preprocessing.

Since it is common to find typos on the web, we thought it would be useful to find out the number of misspelled words in our dataset. The unique misspelled words turned out to be over 50 percent of the words (65 percent) As shown in Figure 5. This may be due to the fact that people can over-exaggerate by repeating letters resulting in linguistically incorrect words. Thus, we are considering eliminating misspelled words from our dataset so as not to confuse the model.

References

[1] Brad. https://github.com/elnagara/BRAD-Arabic-Dataset.

[2] Dataset 100k arabic reviews. https://www.kaggle.com/datasets/abedkhooli/arabic-100k-reviews.

[3] Elements of arabic script. https://commons.wikimedia.org/wiki/File:Elements_of_Arabic_script_improved.png.

[4] Hard. https://github.com/elnagara/HARD-Arabic-Dataset.